25/T26 Suppl 2

FAO Fisheries Technical Paper, No. 26 Suppl. 2
(Distribution restricted)

FRs/T26 Suppl. 2
MFS - Sampling

LIBRARY
NOV 19 1968
FISH & WILDLIFE SERVICE

MANUAL OF SAMPLING AND STATISTICAL METHODS FOR FISHERIES BIOLOGY

Part II - STATISTICAL METHODS

Chapter 6: Continuous Distributions

by
W.B. Hall

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS
ROME, 1968

DOCUMENTS OF THE FISHERY RESOURCES AND EXPLOITATION DIVISION OF FAO DEPARTMENT OF FISHERIES

Documents which are not official FAO publications are issued in several series. They are given a restricted distribution and this fact should be indicated if they are cited. Most of them are prepared as working papers for meetings, or are summaries of information for use of member governments, organizations, and specialists concerned.

DOCUMENTS DE LA DIVISION DES RESSOURCES ET DE L'EXPLOITATION DES PÊCHES DU DÉPARTEMENT DES PÊCHES DE LA FAO

Des documents qui ne figurent pas parmi les publications officielles de la FAO sont publiés dans diverses séries. Ils font seulement l'objet d'une distribution restreinte, aussi convient-il de le préciser lorsque ces documents sont cités. Il s'agit le plus souvent de documents de travail préparés pour des réunions, ou de résumés d'information à l'intention des gouvernements des pays membres, ainsi que des organisations et spécialistes intéressés. Ces séries sont les suivantes:

DOCUMENTOS DE LA DIRECCION DE RECURSOS PESQUEROS Y EXPLOTACION DEL DEPARTAMENTO DE PESCA DE LA FAO

Esta Subdirección publica varias series de documentos que no pueden considerarse como publicaciones oficiales de la FAO. Todos ellos tienen distribución limitada, circunstancia que debe indicarse en el caso de ser citados. La mayoría de los títulos que figuran en dichas series son documentos de trabajo preparados para reuniones o resúmenes de información destinados a los estados miembros, organizaciones y especialistas interesados.

FAO Fisheries Report	**FR /R (No.)**
FAO Fisheries Circular	**FR /C (No.)**
FAO Fisheries Synopsis	**FR /S (No.)**

Special groups of synopses are identified by symbols followed by classification numbers based on indexed code of "Current Bibliography":

SAST Data concerning certain species and fish stocks.

MAST Information on methods and subjects.

OT Oceanographic data.

IT Limnological data.

and

CART Information concerning fisheries and resources of certain countries and regions (FID/S).

Des catégories spéciales de synopses sont identifiées à l'aide de symboles suivis des chiffres de classification basés sur le code d'indexation de la « Current Bibliography »:

SAST Données sur certaines espèces et populations de poissons.

MAST Renseignements sur des méthodes et des sujets.

OT Données océanographiques.

IT Données limnologiques.

et

CART Renseignements sur les pêcheries et les ressources de certains pays et régions (FID/S).

Grupos especiales de sinopsis se distinguen con las siglas siguientes, seguidas por números de clasificación que se basan en las claves de los índices de la « Current Bibliography ».

SAST Datos relativos a ciertas especies y poblaciones.

MAST Sinopsis sobre métodos y materias.

OT Sinopsis sobre oceanografía.

IT Sinopsis sobre limnología.

y

CART Información sobre los recursos acuáticos vivos de algunos países y regiones (FID/S).

FAO Fisheries Technical Paper **FR /T (No.)**

Special groups of Technical Papers are identified by:

RE Indexed lists of experts and institutions drawn from Registers maintained by the Fishery Resources and Exploitation Division.

CB Lists of periodicals, special sections of "Current Bibliography for Aquatic Sciences and Fisheries," special bibliographies and papers concerning documentation problems.

MFS Provisional editions of "FAO Manuals in Fisheries Science."

Some documents also have another identification, if, for example, they have been contributed to a meeting for which papers have been numbered according to another system.

Des catégories spéciales de documents techniques sont identifiées à l'aide des symboles suivants:

RE Listes indexées d'experts et institutions tirées des registres tenus à jour par la Division des ressources et de l'exploitation des pêches.

CB Listes de périodiques, des sections spéciales de la « Current Bibliography for Aquatic Sciences and Fisheries », des bibliographies particulières et des articles sur les problèmes de documentation.

MFS Editions provisoires des « Manuels FAO de science halieutique ».

Il arrive que certains documents portent d'autres numéros d'identification, par exemple, s'ils ont été préparés pour une réunion dont les documents ont été marqués à l'aide d'un autre système.

Grupos especiales de documentos técnicos se identifican por las siglas siguientes:

RE Listas índices de expertos y de instituciones tomadas de los registros que se llevan en la Dirección de Recursos Pesqueros y Explotación.

CB Listas de periódicos, secciones especiales de la « Current Bibliography for Aquatic Sciences and Fisheries », bibliografías especiales y trabajos relativos a los problemas de documentación.

MFS Ediciones provisionales de los « Manuales de la FAO de Ciencias Pesqueras ».

Algunos documentos tienen también otra identificación si, por ejemplo, son contribuciones a una reunión cuyos documentos han sido marcados con arreglo a otros sistemas.

MANUAL OF SAMPLING AND STATISTICAL METHODS FOR FISHERIES BIOLOGY

Part II - STATISTICAL METHODS

Chapter 6

CONTINUOUS DISTRIBUTIONS

by

W.B. Hall
(Aberdeen, Scotland)

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS
Rome, July 1968

PREPARATION OF THIS PAPER

At the Symposium of ICES on Measurement of the abundance of fish populations, held at Madrid in 1963, the need was suggested for a series of "statistical leaflets" each of which would briefly describe a method useful in fisheries research, illustrated where possible by worked examples from fisheries biological studies. ICES subsequently requested FAO cooperation in this project, and a joint Editorial Group was established, composed of J.A. Pope (Aberdeen, Scotland), Convener K.P. Andersen (Copenhagen, Denmark), J.A. Gulland (FAO) and S.J. Holt (FAO). It was decided that a suitable publication vehicle for the "leaflets" would be the existing series FAO Manuals in Fisheries Science, and that each could conveniently be issued as a chapter in No. 3 of that series, the first four chapters of which, written by J.A. Gulland - a member of the Editorial Group - dealt with sampling methods.

The Editorial Group subsequently drew up a plan for the first few statistical chapters and enlisted the help of other scientists to prepare them. The plan is outlined in FAO internal report FB/R85 (October 1964) and in a report to the Comparative Fishing Committee of ICES made by J.A. Pope in Copenhagen, September 1964.

The papers are initially issued as provisional versions in the original language in the series FAO Fisheries Technical Papers, and distributed more widely for comment, and trial use in Training Courses etc. Finally, after further revision they will be published in English, French and Spanish in the "Manuals" series.

PREPARATION DE CE DOCUMENT

Durant le Colloque du CIEM sur l'Evaluation de l'abondance des populations de poissons, tenu à Madrid en 1963, il a été fait mention de la nécessité de créer une série de "Brochures Statistiques" dont chacune décrirait brièvement une méthode utile pour la recherche halieutique. Ces brochures pourraient être illustrées par des exemples pratiques tirés d'études de biologie halieutique. Le CIEM, en conséquence, a sollicité l'aide de la FAO pour ce projet et un Groupe Editorial mixte a été établi, comprenant J.A. Pope (Aberdeen, Ecosse), K.P. Andersen, Organisateur (Copenhague, Danemark), J.A. Gulland (FAO) et S.J. Holt (FAO). Il a été décidé que ces "brochures" seraient à leur place dans la série déjà existante du FAO Manuals in Fisheries Science et que chacune de ces brochures constituerait un chapitre du No. 3 de cette série. J.A. Gulland - membre du Groupe Editorial - serait chargé de la rédaction des quatre premiers chapitres qui traiteraient des méthodes d'échantillonnage.

Le Groupe Editorial a ensuite établi un plan relatif aux premiers chapitres consacrés aux statistiques et s'est assuré le concours d'autres chercheurs qui aideront à leur préparation. Le programme est exposé dans le rapport interne de la FAO, FB/R85 (octobre 1964) et dans un rapport présenté par J.A. Pope au Comité comparatif des Pêches du CIEM, à Copenhague en septembre 1964.

Les documents seront publiés initialement dans leur langue d'origine dans la série des FAO Fisheries Technical Papers et feront l'objet d'une diffusion plus large, pour commentaire. En outre, ils seront utilisés à titre d'essai, notamment dans des cours de formation. Après révision ultérieure, ils seront publiés en anglais, en français et en espagnol dans la série des "Manuals".

PREPARACION DE ESTE DOCUMENTO

En el Simposio del CIEM sobre Medición de la abundancia de las poblaciones de peces, celebrado en Madrid en 1963, fué sugerida la necesidad de publicar una serie de "folletos estadísticos" cada uno de los cuales describiría brevemente un método válido para la investigación pesquera, ilustrado, cuando fuera posible, con ejemplos prácticos según la experiencia de organismos biológicos pesqueros. El CIEM solicitó posteriormente la colaboración de la FAO en este proyecto y para ello se creó un Grupo Editorial conjunto, compuesto de J.A. Pope (Aberdeen, Escocia), convocador K.P. Andersen (Copenhage, Dinamarca), J.A. Gulland (FAO) y S.J. Holt (FAO). Se dicidío que un medio conveniente para la publicación de estos

"folletos" podría ser la ya existente serie de FAO Manuals in Fisheries Science y que cada uno podría publicarse convenientemente como un capítulo en el No. 3 de dicha serie, estando los primeros cuatro capítulos, escritos por J.A. Gulland - miembro del Grupo Editorial - dedicados a los métodos de muestreo.

El Grupo Editorial preparó posteriormente un plan para los primeros capítulos estadísticos y buscó la ayuda de otros científicos para prepararlos. El plan se describe en un informe interno de la FAO, FE/R85 (Octubre 1964) y en el Informe al Comité de Pesca Comparada del CIEM hecho por J.A. Pope en Copenhage, en Septiembre de 1964.

Los documentos se publican como versiones provisionales en el idioma original de la serie de FAO Fisheries Technical Papers y se distribuyen ampliamente para obtener comentarios y para su empleo experimental en Centros de Capacitación etc. Por último, después de su revisión son publicados en inglés, francés y español en la serie de "Manuals".

Distribution

Author
FAO Department of Fisheries
FAO Regional Fisheries Officers
FAO/EPTA Experts in Fisheries Biology
Biological Staff of UNSF Fisheries Projects
Selector SM
ICES

"Current Bibliography" entry

Hall, W.B. (1968) 14-7B008

FAO Fish.tech.Pap., (26)Suppl.2:36 p.
Manual of sampling and statistical methods for fisheries biology. Part 2. Statistical methods. Chapter No. 6. Continuous distributions

Mathematical techniques. Statistical techniques. Frequency distribution. Confidence limits. Curve fitting. Normal distribution.
Co 9-11020.

C O N T E N T S

Chapter	6	CONTINUOUS DISTRIBUTIONS
	6.1	Frequency distributions
	6.2	Measures of location and dispersion
	6.3	Higher moments
	6.4	The Normal distribution
	6.5	Estimating the true mean and variance of a population
	6.6	Fitting a Normal distribution
	6.7	Graphical methods. Uses of arithmetical probability paper
	6.8	The folded Normal distribution
	6.9	Other measures of location and dispersion
	6.10	The bivariate Normal distribution
	6.11	Distributions other than the Normal

6. CONTINUOUS DISTRIBUTIONS by W. B. Hall

This chapter is concerned with sets of measurements which have been summarised into the form of frequency distributions. Numerical and graphical methods of typifying these distributions are discussed with particular reference to the measures of location and dispersion of the Normal distribution.

6.1. Frequency distributions

When a large number of observations is taken of some characteristic such as the lengths of fish or the numbers of lobsters caught by various creels, it is often convenient to summarise these observations in a frequency distribution. This summary makes an appreciation of the main features of the data considerably easier, and also greatly reduces the numerical work involved in calculating measures to describe the data.

At this stage we distinguish between two types of numerical values, viz. continuous and discontinuous (or discrete) variates. If observations are in the form of counts, the variate is discontinuous, i.e. it assumes only the integral values 0, 1, 2 etc. and grouping of this type of variate is usually natural. When data are in the form of measurements they are capable of assuming any value of a continuous range, e.g. 12.4, 14.75, 17.873 etc. although the actual values can be recorded only to a finite number of decimal places because of the limited accuracy of the recording devices employed. In the case of continuous variates any grouping employed is not so natural as with discrete values. The boundaries of the class intervals are arbitrary, being determined by the centres of the intervals and by the nature of the units of measurement used. Initially data are usually recorded in some convenient unit, e.g. centimetres, inches, kilograms and grouping becomes necessary if the number of classes in this unit is too unwieldy. Usually, grouping the data into between 10 and 20 classes of equal size will be found to be suitable remembering that too few classes mean that information is being lost, while too many classes mean that the advantages of grouping may be lost. The width of class intervals should be less than half the standard deviation, otherwise the grouping of the data results in a considerable loss of efficiency in the estimates obtained from the summarised data.

Having decided on the number of classes, the class interval is then defined. For example, if the range of observations is from 10 to 40 cm, a convenient class interval may be 2 cm, and the class intervals may be defined as say 10.00-11.99, 12.00-13.99, 14.00-15.99, if measurements are made to one decimal place, the mid points of these classes being 11.0, 13.0, 15.0 respectively. It is most important that the class interval be defined precisely so that there is no ambiguity as to which class an individual falls in, and that the mid value given to the class is the correct one for

that interval.

Sometimes, however, grouping the data into equal class intervals may destroy the essential features of a distribution. For example, the lengths of several year classes of haddock taken together may range from say 20 to 70 cm. If, as frequently occurs, the youngest year class is predominant, the large majority of fish may be within the range 20-30 cm. Grouping the lengths in 3 cm intervals throughout the whole range would mean that most would fall in the first three classes and the remaining classes would contain relatively few fish. In this case it would be preferable to use a smaller grouping of possibly 1 cm width in the range 20 to 30 cm and a wider grouping of say 5 cm over the rest of the range.

Having summarised the data into a frequency distribution the main features are most clearly seen if presented graphically either as a histogram or as a frequency polygon. In drawing a histogram the character measured is represented along the horizontal axis, and the frequency per class is represented by a rectangle whose base is the range of the class interval and whose height is such as to make the area (not the height) proportional to the number of individuals observed in that class. Class intervals are usually, though not necessarily always, equal. In a frequency polygon the blocks or rectangles of the histogram are replaced by points at the mid points of the class intervals which represent frequencies. These points are then joined together to form a polygon. Usually these points are drawn to make height equal to frequency in which case the class intervals must be of equal length. Figs. 1 and 2 give an example of a histogram and a frequency polygon, the histogram being used here to illustrate continuous variates and the frequency polygon for discrete data. Fig. 3 also illustrates a histogram with unequal class intervals, the tail regions of the distribution being grouped into 5 cm classes.

Sometimes one is interested in the frequency with which a variate takes values equal to or less than some particular value, and this may be seen most easily from a cumulative frequency diagram. This diagram is found by accumulating the frequencies throughout the series and representing these as showing the total frequency to that point. For example the histogram data for Fig. 1 are shown in Fig. 4 as a cumulative frequency polygon. It is often useful to show the percentage of the total frequency on a scale along with the frequency scale as shown in Fig. 4

6.2. Measures of location and dispersion

Several measures are used to typify a set of observations in frequency distribution form. One set of measures is

a) the mode which is the variate value with maximum frequency

b) the median which is the variate value dividing the total frequency into two equal parts

c) the arithmetic mean which is defined as $m = \sum_{1}^{n} x_i / n$ or, if the observations are grouped in the form of a frequency distribution as

$$\sum_{i=1}^{k} f_i x_i \Big/ \sum_{i=1}^{k} f_i$$

where f_i is the frequency observed in the i^{th} class and x_i is the mid point of the i^{th} class and k is the number of classes. In theoretical statistics the concept of infinite collections of observations (infinite populations) is frequently employed. For such populations the concept of an average value remains and its value is denoted by $\mu = E(x)$ where the expression on the right is read as "expected value of x".

Example 6.2.1

Consider the following length measurements of 2 year old whiting from the North Central area of the North Sea, arranged in a length frequency distribution with class intervals of 1 cm.

Class interval (cm)	Length mid point	Frequency
24.5-25.4	25	5
25.5-26.4	26	17
26.5-27.4	27	27
27.5-28.4	28	31
28.5-29.4	29	43
29.5-30.4	30	33
30.5-31.4	31	25
31.5-32.4	32	23
32.5-33.4	33	15
33.5-34.4	34	15
34.5-35.4	35	6
35.5-36.4	36	2
36.5-37.4	37	1
		243

a) In this distribution the mode is the length where the largest frequency occurs, viz. 29 cm.

b) The median has equal frequencies on both sides of it and in this case is the 122nd value, i.e. it lies somewhere in the class with mid point 29 cm. The median is then calculated as

$$L_m = 28.5 + \frac{122-80}{43} = 29.48 \text{ cm}$$

using linear interpolation, where 28.5 is the lower limit of the class containing the 122nd observation, 122 is the accumulated median frequency, 80 is the accumulated frequency to the end of the 28 cm class and 43 is the frequency in the 29 cm class. In this example the total is an odd number and the 122nd value has 121 values on either side of it. If the total had been even, say 244, the median accumulated frequency would be taken as $122\frac{1}{2}$, and the calculation would proceed as above.

Alternatively the cumulative frequency diagram may be used and the median value read off on the x-axis below the point where the frequency diagram cuts the 50% line.

c) By far the most important measure is the arithmetic mean which is obtained by summing the products of frequencies and variate values and dividing by the total frequency thus

$$(5 \times 25 + 17 \times 26 + \ldots + 1 \times 37)/243 = 7236/243 = 29.78 \text{ cm}$$

In this connection considerable labour can be saved by taking a working mean. This could be taken at length 31 cm and the frequency distribution measured from this new origin would be

x	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
f	5	17	27	31	43	33	25	23	15	15	6	2	1

The arithmetic mean on the original scale would be

$$m = 31 + (5 \times -6 + 17 \times -5 + \ldots + 25 \times 0 + \ldots + 2 \times 5 + 1 \times 6)/243$$

$$= 31 - 297/243 = 29.78 \text{ cm}$$

When a desk calculating machine is employed it is often better to avoid using negative numbers, and in this case the working mean could be taken at length 25 cm. The frequency distribution measured from this origin is then

x	0	1	2	3	4	5	6	7	8	9	10	11	12
f	5	17	27	31	43	33	25	23	15	15	6	2	1

Calculations proceed as formerly giving the arithmetic mean on the original scale as

$$m = 25 + (5 \times 0 + 17 \times 1 + \ldots + 25 \times 6 + \ldots + 1 \times 12)/243$$

$$= 25 + 1161/243 = 29.78 \text{ cm}$$

In addition to these location measures it is necessary to measure the variation of the variates about the mean. Obviously two distributions may have exactly the same mean but be widely different in other characteristics such as clustering around the central value. One such measure of this variation is the range, which is the difference between the highest and lowest variate value. This measure is based on only two of the available observations and ignores the information provided by all the intermediate observations. For this, and other reasons, the range is not normally employed as a measure of variability (but see section 6.9).

The measure of dispersion most frequently used is the variance, defined for a finite population of n values as

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

An alternative expression for σ^2 is var(x). The variance then is simply the average value of the squares of the deviations of the observations from their mean value. (The term mean square error was used for the variance in older literature.) When n is infinite the variance is denoted symbolically by $\sigma^2 = E\{(x - m)^2\}$, this expression simply defining the variance as the average (or

expected) value of the squared deviations of the x's from their mean. In practice, we are very largely concerned with estimating the value of the variance from a finite sample of observations.

A computationally simpler version of the variance formula given above is

$$\sigma^2 = \left[\sum x^2 - (\sum x)^2/n\right]/n$$

as with most calculating machines it is possible to obtain the sum $(\sum x)$ and sum of squares $(\sum x^2)$ of a series of observations at the same time.

The square root of σ^2 is then an average deviation from the mean and is known as the <u>standard deviation</u>.

The variance of the whiting lengths given in Example 6.2.1 is thus calculated as follows

$$n = \sum f_i = 243$$
$$\sum f_i x_i = 7236$$
$$\sum f_i x_i^2 = 217030$$
$$(\sum f_i x_i)^2/n = 215472$$
$$S(f_i x_i^2) = 1558 \qquad \text{where } S(f_i x_i^2) = \sum_i^k f_i (x_i - m)^2$$

and $$\sigma^2 = \frac{1558}{243} = 6.41.$$

The standard deviation $\sigma = \pm\sqrt{6.41} = \pm 2.53$

The variance of a series of observations is unchanged if a constant is added to each term, i.e. var(x + c) = var(x). Thus, if the above calculations are repeated on the frequency distribution where the variates have been transformed to the new origin at 31 cm we obtain

$$n = 243, \sum f_i x_i = -297, \sum f_i x_i^2 = 1921, S(f_i x_i^2) = 1558$$

and $\sigma^2 = 6.41$, as before.

Example 6.2.2

Calculate the mean and variance of the following set of carapace lengths of Norway lobsters grouped in 5 mm intervals a) using the variate values recorded and b) using a working mean of 35 mm on a transformed scale.

Carapace length mm	x	15	20	25	30	35	40	45	50	55
Frequency	f	9	588	1235	1210	871	460	192	86	12

a) $n = 4663$

$\sum f_i x_i = 141555 \qquad m_x = 30.36$

$\sum f_i x_i^2 = 4541175$

$S(f_i x_i^2) = 243980 \qquad \sigma^2 = 52.32$

b) If we take a working mean of 35 mm and transform the variates to this new origin and also to a new scale by the transformation $y = \frac{x - 35}{5}$ i.e. working in 5 mm units centred at 35 mm we obtain the following distribution

y	-4	-3	-2	-1	0	1	2	3	4
f	9	588	1235	1210	871	460	192	86	12

and the calculations are

$$n = 4663$$
$$\sum f_i y_i = -4330$$
$$\sum f_i y_i^2 = 13780$$
$$S f_i y_i^2 = 9759.22$$

The mean m_y (say) of this distribution is -0.929 and the mean m_x (say) on the original scale is

$$m_x = 5\, m_y + 35 = 5(-0.929) + 35 = 30.36$$

The variance of the y distribution is 2.093 a value which is 1/25th of the variance of the x's. This is by virtue of the fact that $var(cx) = c^2 var(x)$ where c is a constant, $^1/_5$ in this example.

In both these examples the variances have not been adjusted to allow for the effect of grouping. This effect is discussed elsewhere in this series (chapter 5, Computations) and it is sufficient here to mention that Sheppard's correction for grouping is used when the distribution is continuous and has high order contact with the variate axis at its ends. In the case of the variance of Example 6.2.1 this correction ($h^2/12$ where h is the class interval width) is subtracted from the variance i.e.

$$\sigma^2 = 6.41 - 1/12 = 6.33$$

6.3. Higher moments*

The mean and variance are referred to as the first and second moments of a distribution. Higher moments may be defined; the rth moment of a finite population about zero is $m'_r = \sum f_i x_i^r / n$ and the rth moment about the mean of the distribution is $m_r = \sum f_i (x_i - m)^r / n$. For infinite populations these are replaced by $\mu'_r = E(x^r)$ and $\mu_r = E(x - \mu)^r$.

The moments about zero and about the mean are related, the formulae connecting the first four being:-

$$m_1 = m'_1 - m'_1 = 0; \quad m_2 = m'_2 - (m'_1)^2$$
$$m_3 = m'_3 - 3m'_1 m_2 + 2(m'_1)^3$$
$$m_4 = m'_4 - 4m'_1 m'_3 + 6m'^2_1 m'_2 - 3m'^4_1$$

*This section may be omitted at a first reading.

The general formula expressing m_r in terms of m'_r, m'_{r-1} .... m'_1 is

$$m_r = \sum_{j=o}^{r} \binom{r}{j} m'_{r-j} (-m'_1)^j$$

where $\binom{r}{j} = \frac{r!}{j!\,(r-j)!}$ and $r! = r(r-1)(r-2)$ .... $(3).(2).(1)$. The same expressions hold between μ_r and μ'_r.

The value of these moments is mainly to define further the general shape of a frequency distribution. Clearly the moments of odd order are zero for symmetrical distributions, since for any positive deviation $(x - m)$ there is a corresponding negative deviation $(x' - m) = -(x - m)$ and if r is odd $(x' - m)^r = -(x - m)^r$. If the distribution is not symmetrical the moments of odd order will not vanish and a measure of this asymmetry or skewness is $\beta_1 = m_3^2/m_2^3$ whose square root $\sqrt{\beta_1}$ is zero when the distribution is symmetrical, negative when the distribution is negatively skew (i.e. when the mode is to the right of the mean) and positive when the distribution is positively skew (i.e. when the mode is to the left of the mean) as shown in Fig. 5.

A distribution may be symmetrical but be peaked as in Fig. 6(1) or flat as in Fig. 6(2). This shape characteristic is called kurtosis and the measure usually employed to indicate the degree of kurtosis is $\beta_2 = m_4/m_2^2$. For the Normal distribution β_2 is 3. If $(\beta_2 - 3)$ is positive the distribution is leptokurtic (more peaked than the Normal distribution) and if $(\beta_2 - 3)$ is negative the distribution is platykurtic (flatter than the Normal distribution).

The calculation of higher moments such as those employed to measure shape characteristics is facilitated by working with factorial moments. This is illustrated in the following example

Example 6.3.1

x	y=(x-25)	f	S_1	S_2	S_3	S_4	S_5
25	0	5	243=A				
26	1	17	238	1161=B			
27	2	27	221	923	2972=C		
28	3	31	194	702	2049	5208=D	
29	4	43	163	508	1347	3159	6743=E
30	5	33	120	345	839	1812	3584
31	6	25	87	225	494	973	1772
32	7	23	62	138	269	479	799
33	8	15	39	76	131	210	320
34	9	15	24	37	55	79	110
35	10	6	9	13	18	24	31
36	11	2	3	4	5	6	7
37	12	1	1	1	1	1	1

The columns S_1, S_2 ... S_5 are obtained by successively accumulating from the bottom the observations in the previous column, e.g.

$$A = f_0 + f_1 + f_2 + ... + f_{12} \quad B = f_1 + 2f_2 + 3f_3 + ... + 12f_{12}$$

$$C = f_2 + 3f_3 + 6f_4 + ... 66f_{12} \text{ etc.}$$

Moments about zero are obtained from the accumulated sums A ... E in the following way

$$m_1' = B/A = 1161/243$$

$$m_2' = (2C + B)/A = 7105/243$$

$$m_3' = (6D + 6C + B)/A = 50241/243$$

$$m_4' = (24E + 36D + 14C + B)/A = 392089/243$$

The moments about the mean are obtained from the conversion formulae above, giving on the original scale

$m_1 = 25 + 4.777778 = 29.777778$

$m_2 = 6.411520 \times h^2$ where h is the width of the class interval

$= 6.411520$

$m_3 = 5.791511 \times h^3$

$= 5.791511$

$m_4 = 103.632580 \times h^4$

$= 103.632580$

Applying Sheppard's corrections for grouping (see Chapter 5, Computations) these moments become

$$m_1 = 29.7778$$

$$m_2 = 6.411520 - h^2/12 = 6.3282$$

$$m_3 = 5.7915$$

$$m_4 = 103.632580 - \frac{1}{2} m_2 h^2 + \frac{7}{240} h^4$$

$$= 100.4977$$

Thus $\beta_1 = m_3^2/m_2^3 = 0.1324$ and $\beta_2 = m_4/m_2^2 = 2.5096$. From these values of β_1 and β_2 we see that the above distribution is positively skew and slightly platykurtic.

6.4. The Normal distribution

One of the most important distributions in statistical theory is the Normal or Gaussian distribution. Normal distributions (or distributions which are approximately Normal) are frequently met in practice, and all continuous distributions which are not Normal can be made so by a transformation of the observations.

The Normal distribution is continuous and symmetric. The range of values of the variates is from $-\infty$ to $+\infty$ and the probability density function is given by

$$f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx$$

i.e. the probability of a variate lying within the range x to x + dx is given by the above function, where $x - \mu$ is the distance of the observation x from the true mean μ and σ^2 is the variance. It will be seen that this distribution depends only on the values taken by the two parameters μ and σ^2, and once these are known the distribution is completely specified.

By virtue of the symmetry of the distribution the odd moments about the mean are zero. The first four moments are

$$\mu_1 = \mu$$
$$\mu_2 = \sigma^2$$
$$\mu_3 = 0$$
$$\mu_4 = 3\sigma^4$$

The mean, the median and the mode of this distribution are coincident.

The Normal distribution is tabulated in terms of the proportions of observations lying between specified limits, in many tables, e.g. Fisher and Yates' Statistical Tables for Biological, Agricultural and Medical Research and a table of the Normal distribution is given at the end of this chapter. In this table the total area under the Normal curve (i.e. total frequency) is taken as unity and the areas (frequency of observations) between any two variate values are expressed as a proportion of this total. Immediately we can say that the proportion lying between $-\infty$ and $+\infty$ is unity, and the proportion lying to the right (or left) of the mean is 0.5. The proportion contained between two values of the variate depends only on the distance of these two values from the mean when measured in units of the standard deviation (the square root of the variance) i.e. when the deviations from the mean are standardised by the transformation

$$y = \frac{x - \mu}{\sigma}$$

The quantity y is referred to as a Normal deviate (N.D.). This means that the proportion of the total area lying between say $\mu + \sigma$ and $\mu - \sigma$ is always 0.68 irrespective of the values of the mean and standard deviation. The most important proportions commonly used are

a) that the proportion of the area lying between $\mu + 1.96\sigma$ and $\mu - 1.96\sigma$ is 0.95

b) between $\mu + 2.58\sigma$ and $\mu - 2.58\sigma$ is 0.99 and

c) between $\mu + 3.29\sigma$ and $\mu - 3.29\sigma$ is 0.999.

Example 6.4.1

What percentage of 1 year old whiting will escape from a Courlene cod-end of mesh size 75 mm whose selection factor for whiting is 3.6? If the mean mesh size of the cod-end is increased to

85 mm what further percentage will escape from the cod-end? The mean length of the Normally distributed whiting population is 29.5 cm and the standard deviation is ± 2.35 cm.

We assume that cod-end mesh selection is knife-edged, i.e. that all fish below the 50% length will be released and all above this length will be caught.

Now 50% length = Selection Factor x Mean Mesh Size

= 3.6 x 7.5 = 27.0 cm.

The 50% length lies (27.0-29.5)= -2.5 cm away from the mean. Expressing this distance in units of the standard deviation this gives a Normal deviate of -2.5/2.35 =-1.06. Then from the table of the Normal deviate the proportion of the total area lying below the value -1.06 is 0.145, i.e. 14.5% of the 1+ whiting will escape through the 75 mm cod-end.

The 50% length of the 85 mm cod-end = 8.5 x 3.6 = 30.6 cm. Standardising this length we have N.D. = (30.6-29.5)/2.35 = 0.47. The proportion of the Normal distribution lying below 0.47σ units from the mean is 0.681. The additional proportion released by the 85 mm cod-end is 0.681 - 0.145 = 0.536, i.e. 53.6% more whiting would be released by increasing the mesh size by 10 mm.

6.5. Estimating the true mean and variance of a population

Statistical estimation has as its basis the idea of random sampling. In order that a sample be truly representative of the population from which it is drawn it is necessary to draw the sample in such a way that each individual has an equal probability of being selected. This is a strict requirement and requires adherence to quite precise rules. These rules are given in the chapters on sampling in this series.

The alternative to sampling is complete itemisation of the population. This is frequently impossible and often unnecessary and in this section we concern ourselves with the information about a population which may be obtained from a sample. In all that follows we assume that sampling is random.

If a sample of size n is taken from a population with true mean μ and true variance σ^2, the sample arithmetic mean $\bar{x}$ is an unbiased estimate of the true mean, and the sample variance $s'^2 = \frac{1}{n}\sum(x - \bar{x})^2$ is an estimate of the true variance. Actually, it may be proved that s'^2 is an underestimate of σ^2 but for large n (say n > 100) this is relatively unimportant. For smaller sample sizes an unbiased estimate of σ^2 is provided by the quantity

$$s^2 = \frac{1}{n-1}\sum(x - \bar{x})^2$$

Now, if repeated samples of size n are taken from a population, a distribution of sample means will be obtained. Intuitively these sample means would not be as variable as the individual observations; they would be expected to cluster closer to the true mean, and indeed the variance of

the distribution of sample means is σ^2/n. The square root of this quantity, $\sigma/\sqrt{n}$, is defined as the standard error of the mean. In the case of a parent population which is Normally distributed, the sample means are also Normally distributed, but, even if the parent population is not Normally distributed the sample means will tend to follow a Normal distribution, the approximation being better the larger the sample size.

It follows, therefore, from statements a), b) and c) above that if the parent population is Normal, the following probability statements hold exactly.

$$P(\bar{x} - \mu \leq \frac{1.96\sigma}{\sqrt{n}}) = 0.975$$

$$\text{and } P(\bar{x} - \mu \geq \frac{-1.96\sigma}{\sqrt{n}}) = 0.975$$

$$\text{i.e. } P(\bar{x} - \frac{1.96\sigma}{\sqrt{n}} \leq \mu) = 0.975$$

$$\text{and } P(\bar{x} + \frac{1.96\sigma}{\sqrt{n}} \geq \mu) = 0.975$$

Combining these results gives

$$P(\bar{x} - \frac{1.96\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{1.96\sigma}{\sqrt{n}}) = 1 - 2(1-0.975)$$
$$= 0.95$$

That is, the probability of the range $\bar{x} - 1.96\sigma/\sqrt{n}$ to $\bar{x} + 1.96\sigma/\sqrt{n}$ including the true mean is 0.95. These limits are symmetrical Confidence Limits and in this instance are 95% Confidence Limits. Limits of 99% and 99.9% confidence may be set in the same way by appropriate change of the value of the Normal deviate. Thus, 99% and 99.9% limits are given respectively by $\bar{x} \pm 2.58\sigma/\sqrt{n}$ and $\bar{x} \pm 3.29\sigma/\sqrt{n}$. These results will also hold approximately in the case of non-Normal parent populations.

Example 6.5.1

The true mean of a Normally distributed population is 26.78 cm with true variance 3.78. A random sample of 298 individuals is taken. What is the probability that the mean of this sample will be less than 26.42?

The variance of the sample mean is $\sigma^2/n = 3.78/298$. The standard error of the mean is the square-root of this quantity, namely $\pm$ 0.113.

The difference between 26.78 and 26.42 is 0.36 which is 3.19 times the standard error of the mean. From the table of the Normal probability integral a difference as large as this, or larger, would occur with a probability of 0.0007.

Example 6.5.2

A random sample of size 100 is taken from a Normal distribution with variance 4.90. The mean of the sample is found to be 36.70. Set 99% confidence limits to the true mean of the population.

The standard error of the sample mean is $\pm\sqrt{4.90/100} = \pm 0.221$

The Normal deviate corresponding to 99% is 2.58. The limits within which the true mean is expected to fall with 99% confidence are then

$$36.70 \pm (2.58)(0.221) = 36.13 - 37.22$$

Example 6.5.3

A sample of size 100 is taken providing a mean of 163 gm. Previous large scale investigations had shown that individuals of this kind are Normally distributed with true mean 159 gm and variance 20.68. Is the sample value consistent with having come from this parent population?

This example serves to illustrate the basic reasoning underlying the testing of a statistical hypothesis, a subject which will be dealt with much more fully in a later chapter. There are two possibilities in the present situation; either the sample came from a population with true mean 159 gm and variance 20.68 or it came from a population with a different mean. The former possibility is conventionally referred to as the null hypothesis. The test procedure is to evaluate the probability of obtaining a sample with a mean deviating by at least as much as the observed mean does from the mean specified by the null hypothesis. If this probability is small we may <u>either</u> accept the null hypothesis as being the true one, admitting to the fact that our sample is a rather unlikely one <u>or</u> we may declare that our sample is not an unusual one and that therefore the null hypothesis should be rejected in favour of the alternative hypothesis. It will be noted that no categorical statement is made regarding the tenability of one or other of the hypotheses; we merely accept or reject the null hypothesis on the basis of a computed probability and we may or may not be correct in doing so. If the statistical test is efficient, the probability of rejecting the null hypothesis if it is true is small.

Returning to the example, if the true mean is 159 gm (the null hypothesis) then sample means of 100 individuals will have a distribution with true mean 159 gm and true variance 20.68/100 = 0.2068. The probability of a sample mean deviating by at least 4 gm (163-159 gm = 4 gm) from the true mean is, from tables of the Normal deviate, less than 1 in 1,000. This is such an unlikely event that we may reasonably reject the null hypothesis in this case.

Another very common test of significance is one which involves the comparison of two sample means. Before discussing this we summarize some relevant formulae.

1) If x_1 and x_2 are Normally distributed, uncorrelated quantities, their sum $(x_1 + x_2)$ and their difference $(x_1 - x_2)$ are also Normally distributed.

2) If the true mean values of x_1 and x_2 are m_1 and m_2 respectively the mean values of their sum and their difference are $(m_1 + m_2)$ and $(m_1 - m_2)$. (This result holds whether or not x_1 and x_2 are Normally distributed.)

3) If the true variances of x_1 and x_2 are σ_1^2 and σ_2^2 respectively, the variance of their sum is $\sigma_1^2 + \sigma_2^2$ and that of their difference is also $\sigma_1^2 + \sigma_2^2$. It follows from this that if $\bar{x}_1$ and $\bar{x}_2$ are two uncorrelated sample means based on n_1 and n_2 observations respectively the variance of the difference $(\bar{x}_1 - \bar{x}_2)$ is $(\sigma_1^2/n_1) + (\sigma_2^2/n_2)$.

4) If k_1 and k_2 are any constants, the variance of the quantity $k_1x_1 + k_2x_2$ is $k_1^2\sigma_1^2 + k_2^2\sigma_2^2$. The quantity $k_1x_1 + k_2x_2$ is called a <u>linear combination</u> of x_1 and x_2. We may form a linear combination of any number of variates, say $x_1, x_2, \ldots, x_p$. If the linear combination is $k_1x_1 + k_2x_2 + \ldots + k_px_p$, its variance is $k_1^2\sigma_1^2 + k_2^2\sigma_2^2 + \ldots + k_p^2\sigma_p^2$ provided $x_1, x_2, \ldots, x_p$ are uncorrelated. For example, if x_1 has variance 4 and x_2 has variance 8, the variance of $(2x_1 - \frac{1}{2}x_2)$ is $(2)^2.4 + (\frac{1}{2})^2.8 = 18$. Note that this result holds for variances. The standard deviation of a linear combination is, of course, the square root of its variance.

Example 6.5.4

The head girths of two samples of 100 and 150 cod are measured and found to have mean values of 12.43 cm and 13.46 cm respectively. Could the true mean head girths in the populations from which these samples were drawn be the same? We will assume it is known that the true standard deviations in these populations have the same value 2.84.

Here we have $\bar{x}_1 = 12.43$, $\bar{x}_2 = 13.46$, $n_1 = 100$ and $n_2 = 150$. If m_1 and m_2 are the true mean head girths then, from result 2) above, $\bar{x}_1 - \bar{x}_2$ will have true mean $m_1 - m_2$ and true variance

$$\left(\frac{2.84}{100}\right)^2 + \left(\frac{2.84}{150}\right)^2 = 0.1344$$

The hypothesis we wish to test is that $m_1 = m_2$. This is the null hypothesis, the alternative being that $m_1 \neq m_2$. If the null hypothesis holds, $(\bar{x}_1 - \bar{x}_2)$ will be distributed with zero mean and variance 0.1344. Assuming all distributions to be Normal, the probability of observing a value deviating from zero by 1.03 cm or more is less than 0.01. We conclude therefore that the null hypothesis does not hold.

6.6. <u>Fitting a Normal distribution</u>

Consider the question of fitting a Normal distribution to a sample of observations. The true mean μ and true variance σ^2 being unknown have to be estimated from the data, the estimates being

$\bar{x}$ and s^2. The method is illustrated for the frequency distribution given below.

(1) x(midpoint)	(2) Frequency	(3) N.D.	(4) Proportion to right	(5) Difference	(6) Fitted frequency
		$-\infty$	1.000		
70	5			0.007	4.5
		-2.45	0.993		
71	9			0.011	7.1
		-2.09	0.982		
72	16			0.024	15.5
		-1.73	0.958		
73	26			0.045	29.1
		-1.36	0.913		
74	48			0.072	46.5
		-1.00	0.841		
75	65			0.102	65.9
		-0.64	0.739		
76	85			0.129	83.3
		-0.28	0.610		
77	89			0.146	94.3
		0.09	0.464		
78	90			0.138	89.1
		0.45	0.326		
79	75			0.117	75.6
		0.81	0.209		
80	60			0.090	58.1
		1.18	0.119		
81	38			0.057	36.8
		1.54	0.062		
82	21			0.033	21.3
		1.90	0.029		
83	12			0.017	11.0
		2.26	0.012		
84	7			0.012	7.8
		∞	0.000		

1) The mean and the variance of the observed distribution are calculated giving $\bar{x}$ = 77.26 and s^2 = 7.6739 - 0.0833 (Sheppard's correction) = 7.5906. Obtain the standard deviation $s = \sqrt{7.5906} = \pm 2.755$.

2) Form column (3) headed N.D. (Normal deviate) by standardising the variates, i.e. the Normal deviate = $(x - \bar{x})/s$ remembering to use the end points of the class intervals, e.g. the second value in this column (-2.45) is obtained from (70.5-77.26)/2.755. The distance from the last end points

to the mean is taken as being infinite and the first and last values in this column are then $-\infty$ and $+\infty$.

3) From the table of the Normal Probability Integral find the proportion of the distribution which lies to the right of the appropriate end point. This column (4) is headed "Proportion to the Right".

4) The proportion in a single class is then given by the difference between successive accumulated proportions, i.e. take differences as shown in the column (5) headed "Difference". This then represents the proportion of the total area lying within this class.

5) Multiply each proportion by $\sum f_i = 646$ to form the expected frequencies as shown in the last column (6).

6) If the expected or fitted frequencies are compared with the observed frequencies it is seen that there is quite good agreement between the two sets of values. The difference between the expected frequencies and the observed frequencies can be tested statistically by means of a test known as the χ^2 goodness of fit test. This may be done by obtaining the sum $\sum \frac{(O-E)^2}{E} = \chi^2$ where O is the observed frequency and E is the fitted frequency, and this sum is tested against tabulated values of χ^2 with k-3 degrees of freedom, where k is the number of classes in the distribution. If the fitted frequencies differ appreciably from the observed frequencies this will result in a large value of χ^2. The fit may be regarded as unsatisfactory if the calculated value of χ^2 exceeds the tabulated χ^2 value at the 5% level of probability. Computationally it is easier to work with the χ^2 formula in the form $\sum \frac{O^2}{E} - N$ where N is the size of the sample. In this example $\chi^2 = 647.696-646 = 1.696$ (d.f. = 12) which is not statistically significant, showing that the fitted distribution agrees well with the observed distribution, and that the observed distribution is adequately fitted by a Normal curve.

6.7. Graphical methods. Uses of arithmetical probability paper

Special paper called arithmetical probability paper may be used to provide estimates of the parameters of the Normal distribution by graphical methods. Arithmetic probability paper has been designed to linearise the accumulated Normal frequency distribution.

Two methods of using this graph paper are outlined. The first use is to provide estimates of the mean and standard deviation of a Normal distribution. For this purpose we use the data given in section 6.6.

x mid-point	f	accumulated f.	acc. f %
70	5	5	0.8
71	9	14	2.2
72	16	30	4.6
73	26	56	8.7
74	48	104	16.1
75	65	169	26.2
76	85	254	39.3
77	89	343	53.1
78	90	433	67.0
79	75	508	78.6
80	60	568	87.9
81	38	606	93.8
82	21	627	97.1
83	12	639	98.9
84	7	646	100.0

The accumulated frequencies, expressed as percentages of the total frequency, are shown in the final column above. These percentages are plotted against the end points of the class intervals to which they refer. Apart from sampling variation, these points will, if the data are Normally distributed, fall on a straight line which can be fitted by eye (Fig. 7). The mean is read off on the variate scale where the fitted line crosses the 50% level, in this example at 77.3. The standard deviation of the distribution is estimated using the fact that the area lying within the range $\mu \pm 1.96\sigma$ contains 95% of the total distribution. If we read off the variate values corresponding to 2.5% and 97.5% on the probability scale the standard deviation is then obtained by dividing the difference between these variate values by 2 x 1.96 = 3.92. In this example

$$x_{0.025} = 71.8,\ x_{0.975} = 82.7 \therefore \sigma = \frac{82.7-71.8}{3.92} = \pm 2.78$$

(cf. calculated values $\bar{x} = 77.26$ and $s = \pm 2.76$). Of course any other Normal deviate can be used instead of 1.96, and often a more convenient procedure is to read off the points at 84.1% and 15.9% corresponding to one standard deviation on each side, dividing the difference between $x_{0.84}$ and $x_{0.16}$ by 2.

An extended use of this paper is to provide estimates of the parameters of a mixture of two or more Normal distributions. The method of fitting will be described for a mixed population containing only two Normal distributions.

When a population is a mixture of two Normal distributions the plot of the accumulated frequencies no longer falls on a straight line but on an S-shaped curve. The method of resolution consists of fitting two straight lines such that, on combination, they reproduce the curve. The

method has a number of important reservations. For example, (i) it must be known beforehand that the observed distribution is a mixture of Normal distributions, because many observed individual distributions are skew and will exhibit non-linearity when plotted on arithmetic probability paper, (ii) even if there are two groups, both must be Normally distributed, as departures from this may influence the final result seriously, (iii) extensive data must be used when making the resolution.

The following example containing a mixture of 0+ and 1+ whiting from the West Coast of Scotland illustrates the method of resolving into the two components.

Length (mid-point)	Frequency	Acc. freq.	Acc. freq. %	Adjusted %	Expected frequency
20	4	4	0.89	2.2	4
21	14	18	4.0	10.0	12
22	29	47	10.4	26.0	29
23	41	88	19.6	49.0	46
24	47	135	30.0	75.0	48
25	27	162	36.0	90.0	37
26	32	194	43.1	5.2	29
27	41	235	52.2	20.3	35
28	45	280	62.2	37.0	44
29	50	330	73.3	55.5	51
30	52	382	84.9	74.8	47
31	36	418	92.9	88.2	34
32	17	435	96.7	94.5	20
33	10	445	98.9	98.2	9
34	5	450			5

The accumulated frequencies, expressed as percentages, are plotted against the end-points of the classes as in the previous example (see Fig. 8). These percentages fall on a shallow S-shaped curve. The proportion of mixture is found from the point of inflexion on this curve. This is, of course, a subjective estimate and there is no indication of its precision. Apart from the trivial case where the populations are only slightly mixed, this can cause a major source of inaccuracy in the estimates obtained.

In this example the point of inflexion occurs at 26.1 cm which, from the column of accumulated frequency percentages, corresponds to a value of 40%. This indicates that 40% of the individuals are from the first population and 60% are from the second. As the percentages have been calculated using the total number in the population they are now adjusted to allow for the fact that the first population contains 40% of the total, i.e. all percentages up to the point of inflexion are multiplied by 100/40. The results are shown in the column headed Adjusted %. These points fall on a straight line which is fitted by eye. This is the lower line of Fig. 8. In the case of a

mixture of two populations one may then start at the other end of the distribution and adjust the percentages from that end in exactly the same way. For example, at the lower end of the 34 class interval (33.5 cm) would be plotted the percentage (reading on the top scale of the graph paper) $\frac{5}{450} \times 100 \times \frac{100}{60} = 1.9\%$ and all the remaining points to the point of inflexion would be fitted accordingly.

Alternatively an easier computational method is to adjust the accumulated percentages after the point of inflexion by subtracting the percentage at the point of inflexion from each accumulated percentage and from the denominator in the multiplying term. Thus the first accumulated percentage after the point of inflexion is 43.1. The adjusted percentage this represents in the second population is (43.1-40)(100)/(100-40) = 5.2. The remaining points are calculated in the same way and fall on a straight line which is fitted by eye. The two means of the constituent populations are read from the two fitted lines. The graphical estimates of the parameters of the two populations are then

p (the proportion of mixing) 0.40:0.60, $\therefore$ $n_1 = 180$, $n_2 = 270$

$\bar{x}_1 = 23.5$ $\bar{x}_2 = 29.1$

$s_1 = \pm 1.5$ $s_2 = \pm 2.1$ $\chi^2(df = 9) = 5.81$

This method of resolving a mixture of Normal distributions can be extended to the case of several distributions as shown by Cassie (1954).

Having resolved the mixture into its component distributions these can be fitted using these estimated parameters. The goodness of fit may be tested in the manner illustrated in section 6.6 remembering that three degrees of freedom are lost for each Normal distribution obtained. The value of χ^2 in this example is 5.81 (d.f. = 9).

Various algebraic methods have been used to tackle this problem, but in the main these involve a considerable amount of computation and it is unnecessary to elaborate them within this series. An account of this is given by Rao (1952).

6.8. The folded Normal distribution*

Sometimes measurements which may be positive or negative are taken without their algebraic signs. This can arise for example where duplicate observations are being made, and the difference is always taken as the smaller observation from the larger. The resulting observations are then always positive, the negative differences being now indistinguishable from the positive ones.

If the underlying distribution of the positive and negative values when measured with their correct signs is Normal, the effect of neglecting the sign is to fold the Normal distribution,

*This section may be omitted at a first reading.

adding the negative values to the corresponding positive values. Fig. 9 shows the effect of a fold in a Normal distribution at the point x_o, all values below this point being added to the corresponding values above this point. This fold can occur at any point in the distribution, one point in particular being the mean of the original distribution when the resultant distribution is a half-Normal distribution.

The effect of folding is to make the mean μ_f of the folded distribution larger than the mean μ of the original distribution, and to make the variance σ_f^2 of the folded distribution less than the original variance σ^2.

Writing $\theta = \mu/\sigma$ and $\theta_f = \mu_f/\sigma_f$, estimates of μ and σ^2 are provided by the equations

$$\hat{\mu} = \hat{\theta}\sqrt{\left[(\mu_f^2 + \sigma_f^2)/(1 + \hat{\theta}^2)\right]} \qquad (6.8.1)$$

$$\hat{\sigma}^2 = (\mu_f^2 + \sigma_f^2)/(1 + \hat{\theta}^2) \qquad (6.8.2)$$

where $\hat{\mu}$ is the estimated value of μ, $\hat{\sigma}^2$ the estimated value of σ^2 and $\hat{\theta}$ is $\hat{\mu}/\hat{\sigma}$. The quantity θ^2 is estimated from the quadratic equation in $\hat{\theta}^2$

$$(1 + \hat{\theta}^2)^2 B - (3 + 6\hat{\theta}^2 + \hat{\theta}^4) = 0 \qquad (6.8.3)$$

where $B = \mu_f'(4)/(\mu_f'(2))^2$, $\mu_f'(4)$ and $\mu_f'(2)$ being the fourth and second moments about zero of the folded distribution.

Example 6.8.1

Surface temperature readings by two thermometers were made at stations two miles apart at 166 different positions over the northern North Sea. The differences between readings were taken and recorded only as positive quantities. The data, grouped in classes of 0.02°C, are shown below.

Difference (x)	0.01	0.03	0.05	0.07	0.09	0.11	0.13
Frequency	57	48	31	19	7	3	1

Assuming the underlying distribution of differences to be Normal, estimate its mean and standard deviation.

The first four moments of the folded distribution about zero are

$$\mu_f = 0.03602 \qquad \mu_f'(3) = 0.00013878$$

$$\mu_f'(2) = 0.001984 \qquad \mu_f'(4) = 0.000011286$$

From these σ_f^2 is estimated as 0.0006867 and B as 2.8672. The quadratic equation (6.8.3) is

$$(1 + \hat{\theta}^2)^2 2.8672 - (3 + 6\hat{\theta}^2 + \hat{\theta}^4) = 0$$

which has the solutions $\hat{\theta}^2 = 0.3471$ and -0.2049. Only the solution 0.3471 applies and $\hat{\theta}$ is estimated as 0.5892.

Using (6.8.1) and (6.8.2) we get

$$\hat{\mu} = 0.5892\sqrt{\left\{\frac{(0.03602)^2 + (0.0006867)}{1 + 0.3471}\right\}}$$

$$= 0.0226$$

$$\hat{\sigma} = 0.0384$$

This result means that one thermometer was reading higher than the other by 0.0226°C. The standard deviation of a single thermometer reading is $\sqrt{\frac{(0.0384)^2}{2}} = 0.0272$.

6.9. Other measures of location and dispersion

The measures of location and dispersion which have been dealt with so far are the arithmetic mean, variance and the standard deviation. Other, usually less efficient, measures of location and dispersion are, however, available. Reference has already been made to the median, for example. These other measures are often useful because of their ease and rapidity of calculation and are therefore particularly useful when dealing with large amounts of data.

The most frequently employed alternative measures are those based on order-statistics, the r^{th} order-statistic of a sample of observations being simply the r^{th} smallest variate value. Thus, if the sample observations are 13.1, 9.6, 2.4 and 7.8 the first order-statistic is 2.4, the second is 7.8 and so on.

For symmetrical distributions the median, i.e. the variate value exceeded by half the observations, provides an estimate of the mean. In large samples from a Normal parent population the variance of the median is $1.57\sigma^2/n$ which exceeds the variance of the mean, σ^2/n. This implies that 57% more observations are required to achieve the same accuracy in the median as in the arithmetic mean, or, to put it another way, the median is only $\frac{1}{1.57} \times 100 = 64\%$ as efficient as the arithmetic mean. However, in large samples, the median, which is quickly obtained, may be sufficiently accurate.

Another estimate of the mean is the mid-quartile $\{x_{(0.25n)} + x_{(0.75n)}\}/2$, where $x_{(0.25n)}$ represents, in a fairly obvious notation, the lower quartile. In large samples from a Normal parent population this has a variance of $1.25\sigma^2/n$ and hence an efficiency of 80% relative to the arithmetic mean.

Both the median and mid-quartile are obviously sensitive, as measures of the mean, to asymmetry in the parent distribution. Alternative measures are, therefore, to be preferred when the samples are drawn from skew distributions. Quenouille (1959) has suggested the following estimate of the mean in situations where parent Normality is doubtful

$$\frac{1}{6}\{x_{(0.0625n)} + x_{(0.25n)} + 2x_{(0.5n)} + x_{(0.75n)} + x_{(0.9375n)}\}$$

The variance of this estimate when the parent population is Normal is $1.04\sigma^2/n$. When the parent population is the exponential distribution, a skew distribution, the variance is $1.10\sigma^2/n$. The latter might be employed as a conservative estimate in all situations where extreme skewness is apparent in the data.

Several short-cut methods based on order statistics have been proposed for estimating variability. The measure most widely employed is the sample range. This is the largest observation minus the smallest, or the last order statistic minus the first. In samples from a Normal parent population this quantity estimates a multiple of the standard deviation, the multiple varying with sample size. Thus, if w_n is the range in a sample of size n, then w_n is an estimate of $d_n\sigma$. Hence w_n/d_n is an estimate of σ. Values of d_n and $1/d_n$ are given for a range of sample sizes in "Biometrika Tables for Statisticians" Vol. I (Pearson and Hartley, 1954) and some of these are reproduced below.

n	d_n	$1/d_n$	n	d_n	$1/d_n$
2	1.12838	0.8862	9	2.97003	0.3367
3	1.69257	0.5908	10	3.07751	0.3249
4	2.05875	0.4857	12	3.25846	0.3069
5	2.32593	0.4299	14	3.40676	0.2935
6	2.53441	0.3946	16	3.53198	0.2831
7	2.70436	0.3698	18	3.64006	0.2747
8	2.84720	0.3512	20	3.73495	0.2677

The efficiency of the range as an estimate of the standard deviation decreases with increasing sample size. For this reason it is preferable to estimate the standard deviation by first dividing the total number of available observations into groups, each containing the same number of observations (say 8, 9 or 10) and to average the ranges obtained. Care should be taken to ensure that the observations are randomly allocated to each group.

Example 6.9.1

Using the mean sample range, estimate the standard deviation of the following set of 50 observations grouped into five sets of ten observations each.

(1)	(2)	(3)	(4)	(5)
23	29	21	27	24
29	21	23	24	22
28	25	26	22	29
28	22	21	22	22
28	23	24	23	27
27	28	29	21	26
25	30	30	21	23
27	29	28	29	24
24	28	29	28	22
27	22	26	30	30

The ranges in each group are: (1) 6, (2) 9, (3) 9, (4) 9, (5) 8. Hence $\bar{w}_{10} = 41/5 = 8.2$ and the estimated standard deviation is $(0.3249)(8.2) = 2.7$ (cf. $s = 3.0$).

Apart from its use as an estimate of the standard deviation, the sample range may be used in conjunction with the value of s, calculated in the usual way, as a rough test for the Normality of the observations. Thus, the existence of an extreme observation will have a more marked effect on w_n than on s and the ratio w_n/s will exceed the tabulated value d_n[‡]. On the other hand, if the observations are clumped at the extremes of the range, the value of w_n/s will be less than the tabulated value of d_n.

The range in small samples is fairly insensitive to departures from Normality. As the sample size increases, the range is affected by departures from Normality, especially skewness. An estimate of s which is robust against departures from Normality is provided by

$$s = \tfrac{1}{4}\{y(0.0625n) + \tfrac{3}{4}y(0.25n) - \tfrac{3}{4}y(0.25n) - y(0.9375n)\}$$

This estimate employs the same order-statistics as those used in the estimate of the mean given earlier and both mean and standard deviation may be estimated simultaneously once these order-statistics have been determined.

6.10. The bivariate Normal distribution

So far we have been concerned with the frequency distributions of measurements of only one character of an individual, i.e. we have been considering univariate distributions of say length, weight, etc. Measurements are, however, often taken of two or more characters of an individual and we shall consider here the extension from one character to two characters, i.e. the extension from the univariate case to the bivariate case. Elsewhere in this series multivariate distributions will be considered and this section deals only with some of the basic principles of the univariate Normal distribution extended to the bivariate Normal distribution.

The weights and lengths of 496 two-year-old North Sea herring were taken and grouped into the following two-way table. The weights were grouped in 10 gm intervals and the lengths were grouped in 0.4 cm intervals, the body of the table containing the frequency of occurrence of each joint class.

---

‡More elaborate procedures for the rejection of outlying observations have been devised and will be discussed in another chapter.

Weight/Length Distribution of 2 year old North Sea Herring

Class length	Weight Mid-point	110-119 115	120-129 125	130-139 135	140-149 145	150-159 155	160-169 165	170-179 175	180-189 185	190-199 195	200-209 205	210-219 215	Marginal length distribution
23.5-23.9	23.8	1	4										5
24.0-24.3	24.2	5	2	6	2								15
24.4-24.7	24.6	3	10	15	5	2	1						36
24.8-25.1	25.0	1	2	18	12	7	3						43
25.2-25.5	25.4		1	11	20	18	5	4					59
25.6-25.9	25.8			10	21	17	19	7					74
26.0-26.3	26.2			3	12	23	21	20	7	1			87
26.4-26.7	26.6				3	12	19	22	13	3	1		73
26.8-27.1	27.0					3	8	14	9	4	2		40
27.2-27.5	27.4					1	2	10	5	7	7	4	36
27.6-27.9	27.8								5	5	4	5	19
28.0-28.3	28.2									3	3	3	9
		10	19	63	75	83	78	77	39	23	17	12	496

This distribution can be represented graphically in an analogous form to the diagrammatic representation of the univariate case. Fig. 11 shows a three dimensional histogram, called a stereogram, where a pillar is erected on each cell corresponding to the frequency at each length and weight occurring in that cell. As each cell is of equal area the volume of the pillar then represents the frequency in any particular length-weight class. This figure shows that maximum frequency is concentrated in the central part and tails off to the outer edges. The theoretical bivariate Normal distribution when drawn in continuous form is like a bell.

The marginal length and weight distributions are also shown graphically in Fig. 11. These are of course univariate distributions obtained by summing the frequencies for a given weight or length and are Normal if the bivariate distribution is itself Normal. The bivariate Normal distribution depends on five parameters (cf. the univariate Normal depends on two) μ_1, μ_2, σ_1, σ_2 and ρ where μ_1, μ_2, σ_1, σ_2 are the means and standard deviations of the marginal distributions and ρ, the correlation coefficient, measures the degree of association between the two variates.

If the bell shaped curve is cut parallel to the plane of the axes of the two variates the resulting figure will be an ellipse, and this gives another method of representing a bivariate distribution graphically. Fig. 10 shows the contours of the length-weight distribution corresponding to particular frequencies. Superimposed on these contours are the ellipses of a fitted bivariate Normal distribution at corresponding probability levels. Diagrammatically one can see the divergence of the observed frequency distribution from the theoretical bivariate Normal

distribution; in this particular case the agreement between the observed and the theoretical distributions is quite good, although there is some divergence in the upper tail area.

If the bivariate distribution is cut at right angles to the plane of the variate axes and parallel to one of the axes the resulting curve will be in the univariate Normal form. If the cut is made parallel to say the weight axis the resultant weight frequency distribution for a given length is known as the conditional weight distribution, and similarly if the bivariate distribution is cut parallel to the length axis the resulting length distribution is the conditional length distribution for a given weight. The conditional weight distribution for a given value of $L = L_o$ (say) is Normally distributed with mean $\mu_w + \rho\sigma_w/\sigma_L(L_o - \mu_L)$ and variance $\sigma_w^2(1 - \rho^2)$ and the conditional length distribution for a given weight $W = W_o$ (say) is Normally distributed with mean $\mu_L + \rho\sigma_L/\sigma_w(W_o - \mu_w)$ and variance $\sigma_L^2(1 - \rho^2)$. The means of the conditional distributions lie on the straight lines shown superimposed on Fig. 10. It is seen that the point of intersection of these lines is at the means of the two marginal distributions. (For further details see chapter 7 on Regression and Correlation.)

6.11. Distributions other than the Normal

In this chapter we have been concerned principally with the Normal distribution. However, many other types of distributions occur quite naturally and are in no sense to be thought of as abnormal. The Binomial and Poisson distributions, which have wide applications in statistics, are discrete distributions (i.e. they apply to random variates which may assume only integral values) and may be quite skew. These particular distributions are discussed elsewhere in this series. Again, the distributions of many functions of observations (e.g. the variance) are non-Normal. Statistics is very largely concerned with the distribution of functions of observations and these will likewise be studied in other chapters.

Here we merely note that in situations where the frequency distribution of the parent population is not Normal it may be adequately described by one of a family of curves devised by Karl Pearson and known as Pearson's Curves. This very wide class of frequency curves includes the Normal and several of the more important sampling distributions as particular cases. The reader who wishes a detailed exposition of the Pearson Curves is referred to Elderton (1938), and to Lee (1920) for an example of the fitting of a Pearson Type I Curve to the length distribution of male and female herring.

The use of such curves to describe the frequency distribution of the parent population is, in general, of academic rather than practical relevance since, as already noted, interest centres more on the distributions of functions of sample observations rather than on the distribution of the

parent population itself. All the important sampling distributions have been very thoroughly studied and extensively tabulated and, although based on the assumption of parent Normality, the effects of particular types of departure from Normality have been investigated. In many situations the effect of non-Normality in the parent is small although this is not always true. Again, this will be treated in another chapter.

REFERENCES

Cassie, R.M., 1954 Some uses of probability paper in the analysis of size frequency distributions. Aust.J.mar.Freshwat.Res., 5:513-22

Elderton, Sir W.P., 1938 Frequency curves and correlation. London C.U.P., 3rd ed., 271 p.

Fisher, Sir R.A. and F. Yates, 1963 Statistical tables for biological, agricultural and medical research. Edinburgh, Oliver and Boyd, 6th ed., 146 p.

Lee, R.M., 1920 A review of the methods of age and growth determination in fishes by means of scales. Fishery Invest., Lond.(2), 4(2):32 p.

Pearson, E.S. and H.O. Hartley, 1954 Biometrika tables for statisticians. London C.U.P., 238 p.

Quenouille, M.H., 1959 Rapid statistical calculations. London, Griffin, 81 p.

Rao, C.R., 1952 Advanced statistical methods in biometric research. New York, J. Wiley, 390 p.

FIGURE LEGENDS

Figure 1. Length frequency distribution of a sample of 243 2+ whiting

Figure 2. Frequency polygon of the number of baskets of fish per haul taken in 42 hauls

Figure 3. Histogram of length composition of Barents Sea haddock illustrating grouping in different class widths

Figure 4. Cumulative frequency polygon of data shown in Figure 1

Figure 5. Skewness

Figure 6. Kurtosis

Figure 7. Arithmetical probability paper used to estimate the mean and standard deviation of a normal distribution

Figure 8. Arithmetical probability paper used to resolve a mixture of two normal distributions into the component distributions

Figure 9. The effect of folding a normal distribution at the point x_o

Figure 10. Contour diagram of length weight bivariate distribution

Figure 11. Weight-length stereogram

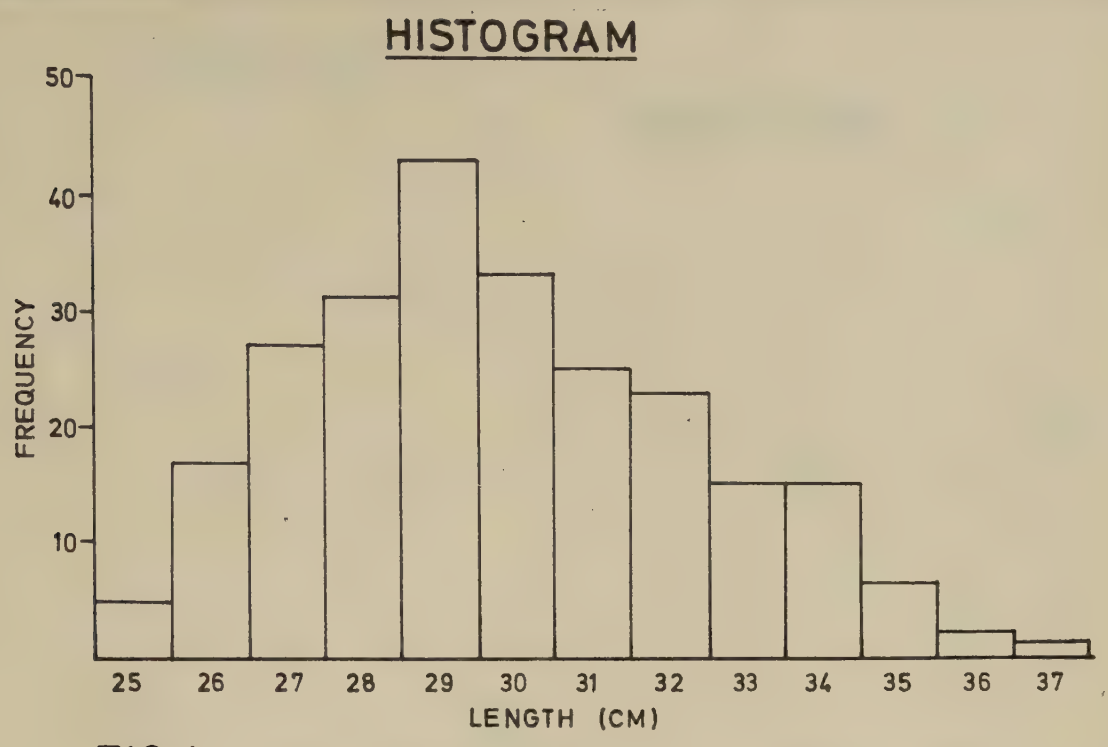

FIG.1 LENGTH FREQUENCY DISTRIBUTION OF A SAMPLE OF 243 2+ WHITING

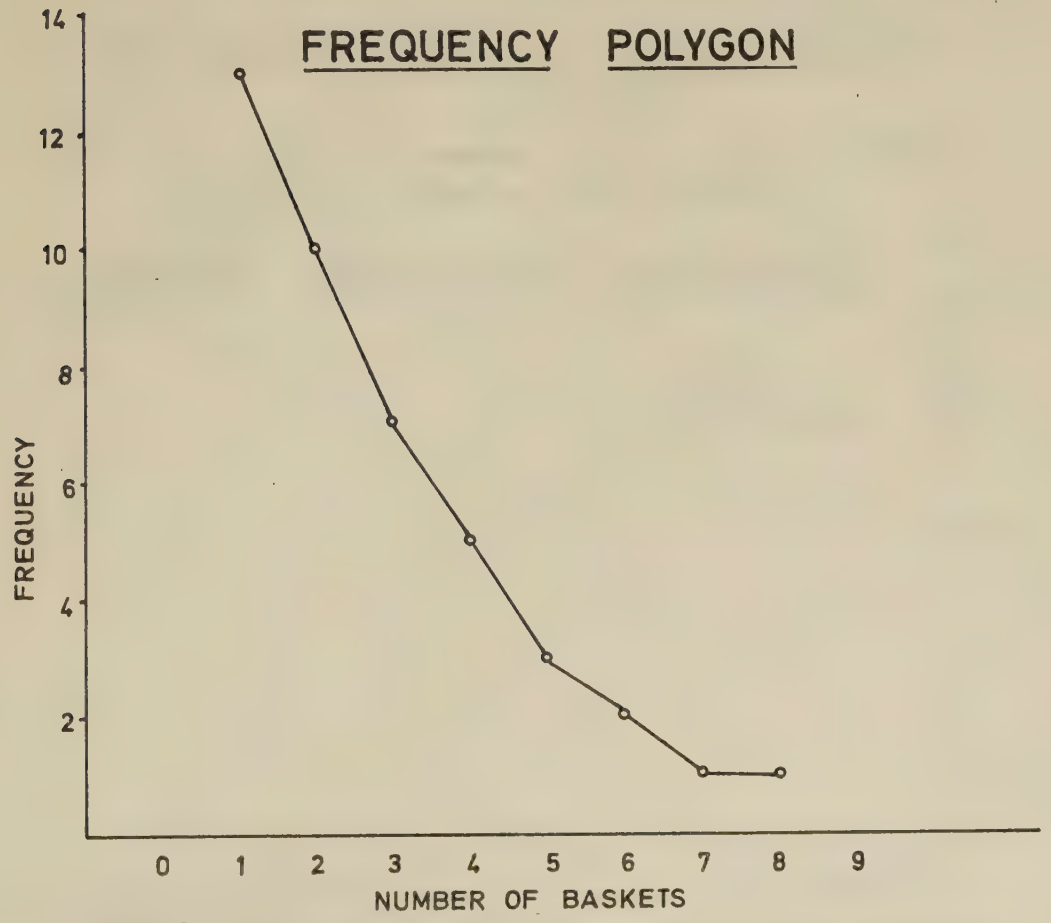

FIG.2 FREQUENCY POLYGON OF THE NUMBER OF BASKETS OF FISH PER HAUL TAKEN IN 42 HAULS.

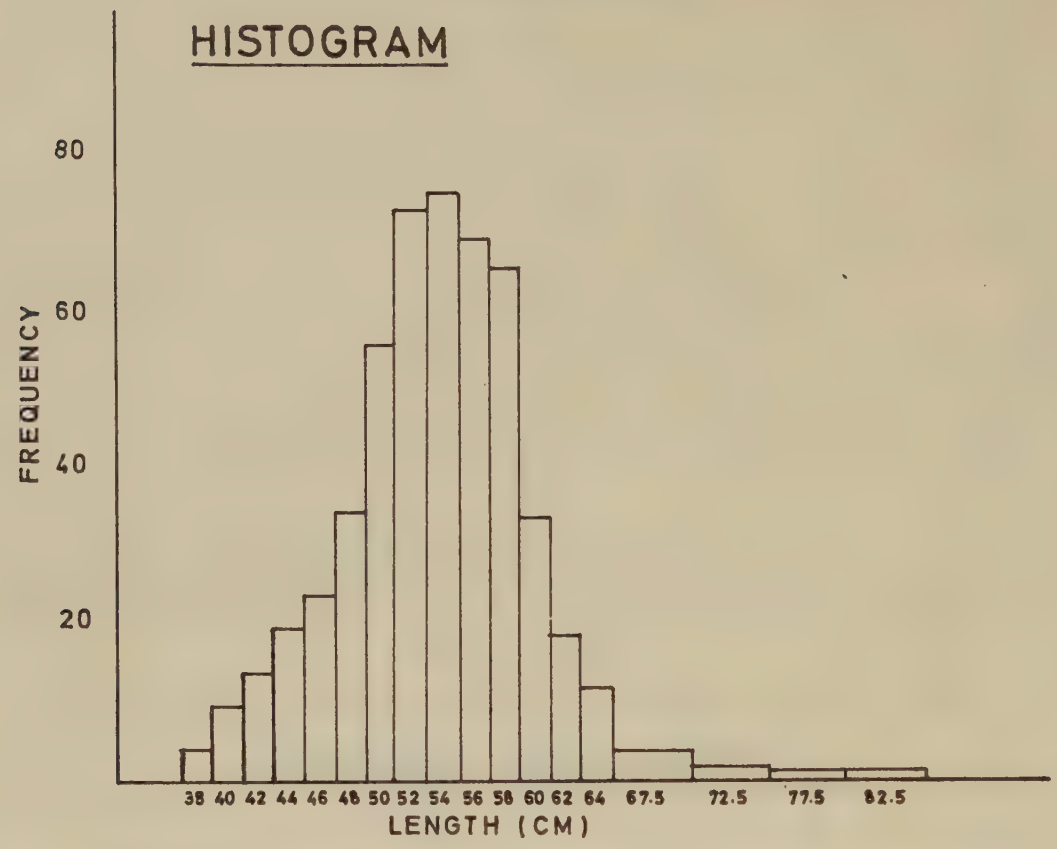

FIG. 3 HISTOGRAM OF LENGTH COMPOSITION OF BARENTS SEA HADDOCK ILLUSTRATING GROUPING IN DIFFERENT CLASS WIDTHS.

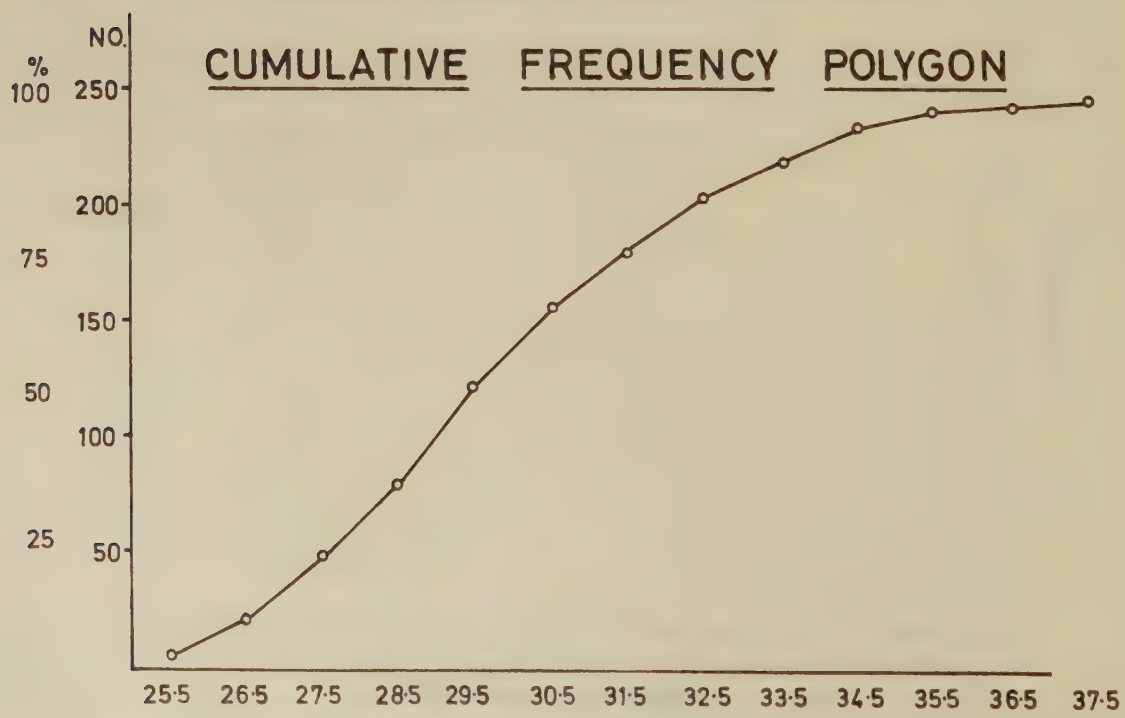

FIG. 4 CUMULATIVE FREQUENCY POLYGON OF DATA SHOWN IN FIG. 1.

SKEWNESS

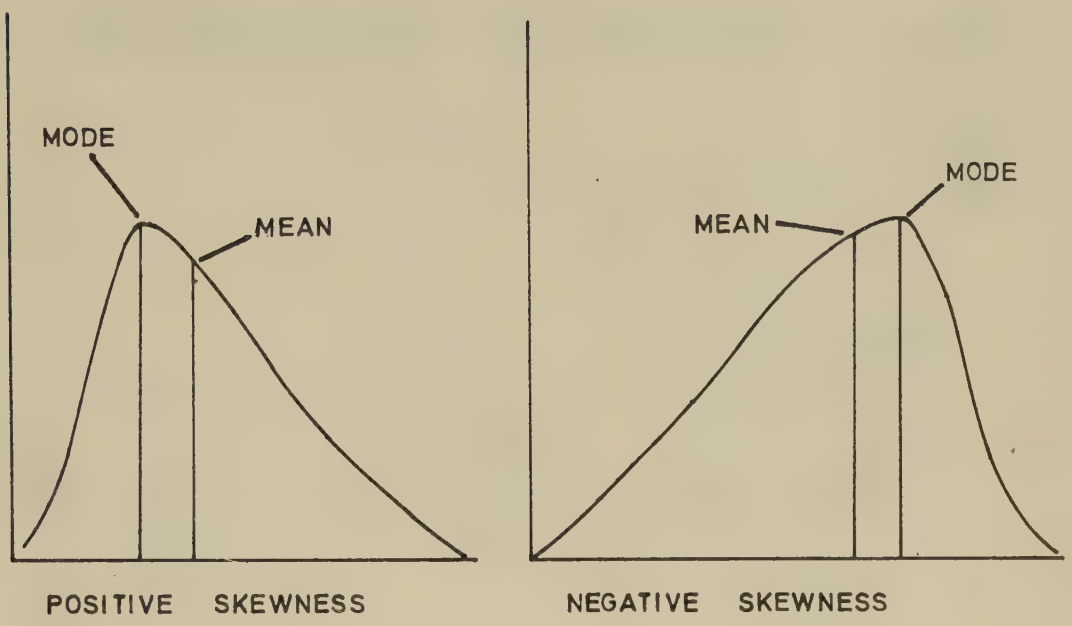

FIG. 5

KURTOSIS

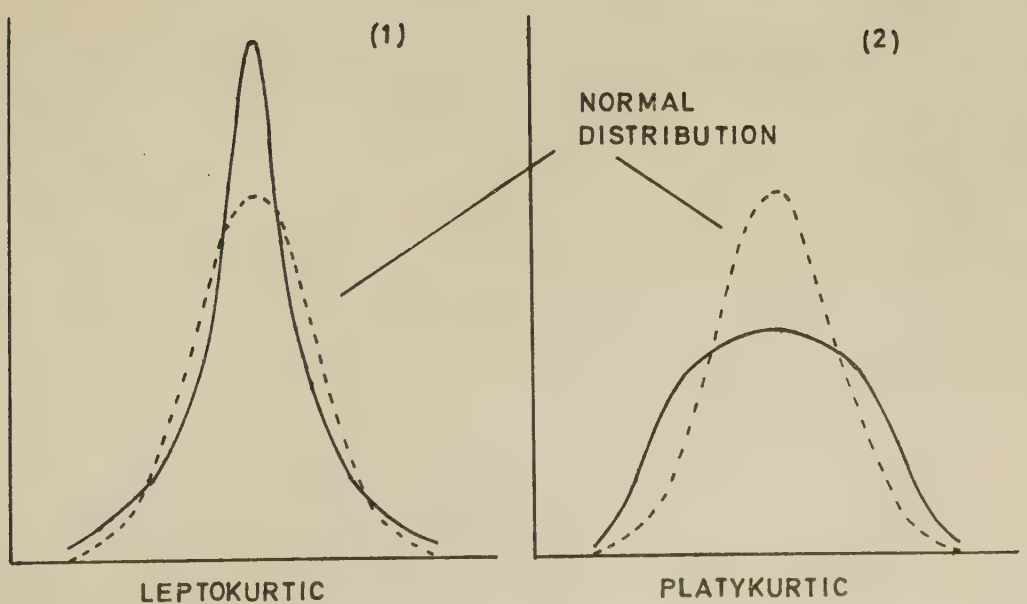

FIG. 6

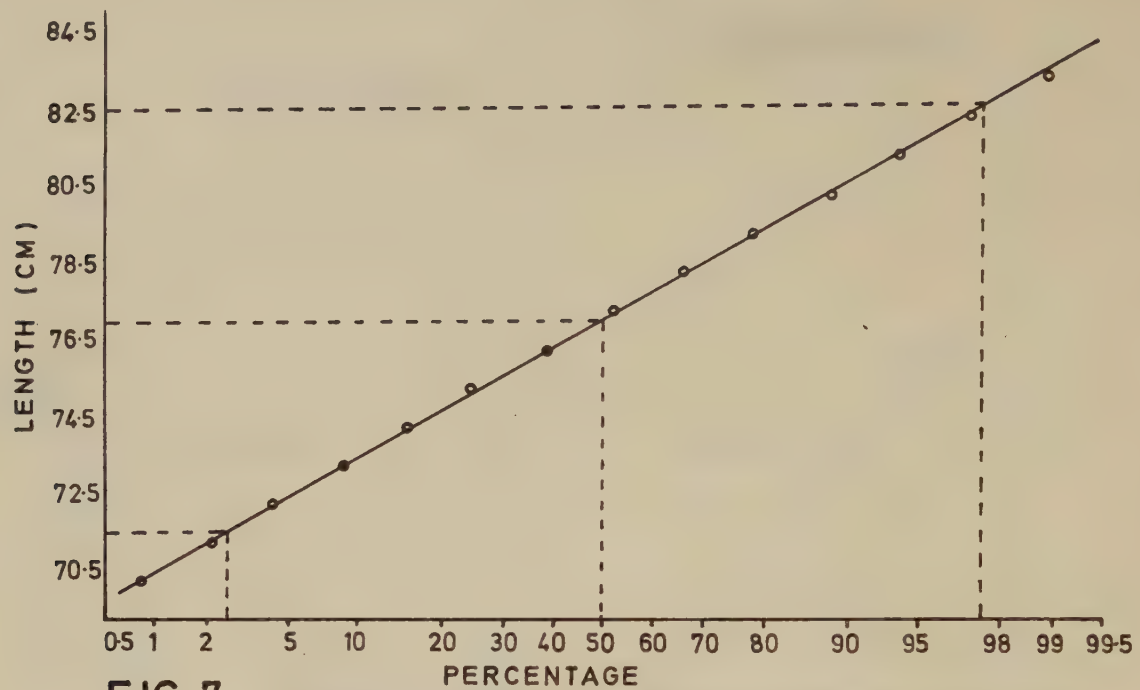

FIG. 7 ARITHMETICAL PROBABILITY PAPER USED TO ESTIMATE THE MEAN AND STANDARD DEVIATION OF A NORMAL DISTRIBUTION .

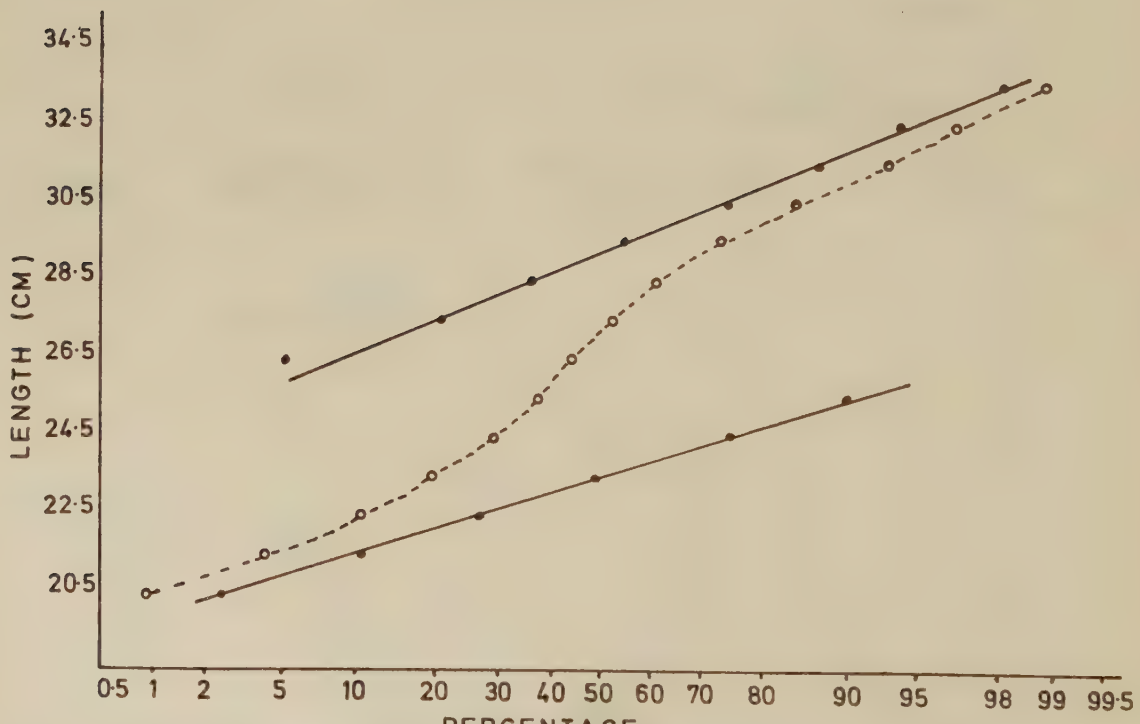

FIG. 8 ARITHMETICAL PROBABILITY PAPER USED TO RESOLVE A MIXTURE OF TWO NORMAL DISTRIBUTIONS INTO THE COMPONENT DISTRIBUTIONS .

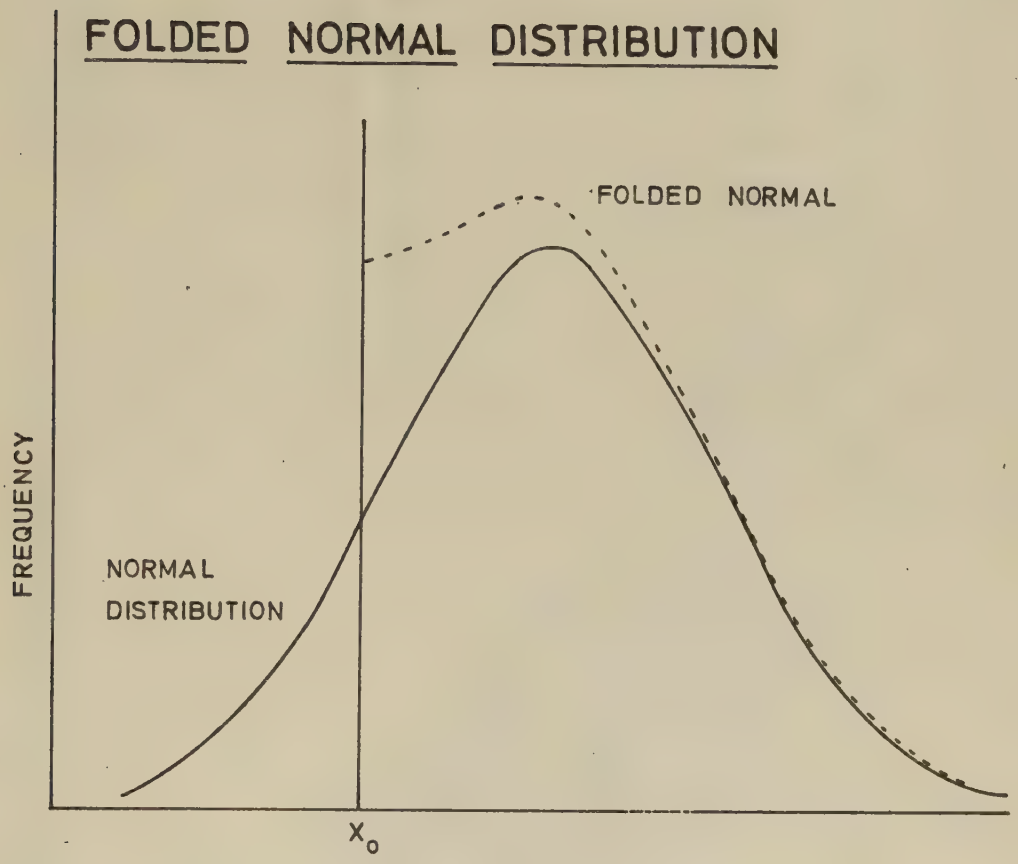

FIG. 9 THE EFFECT OF FOLDING A NORMAL DISTRIBUTION AT THE POINT X_0

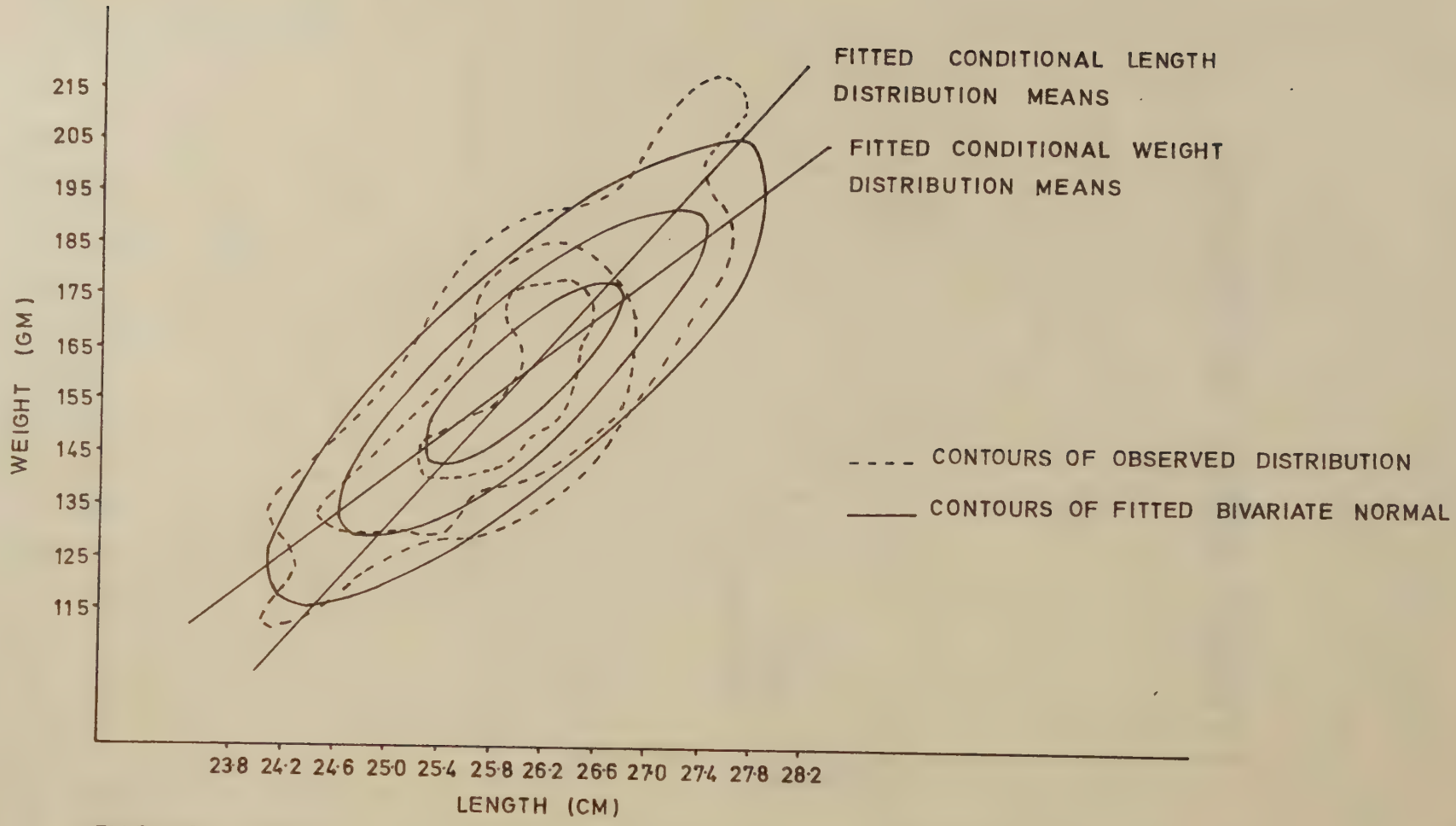

FIG.10 CONTOUR DIAGRAM OF LENGTH WEIGHT BIVARIATE DISTRIBUTION

WEIGHT - LENGTH
STEREOGRAM

2+ HERRING 1963

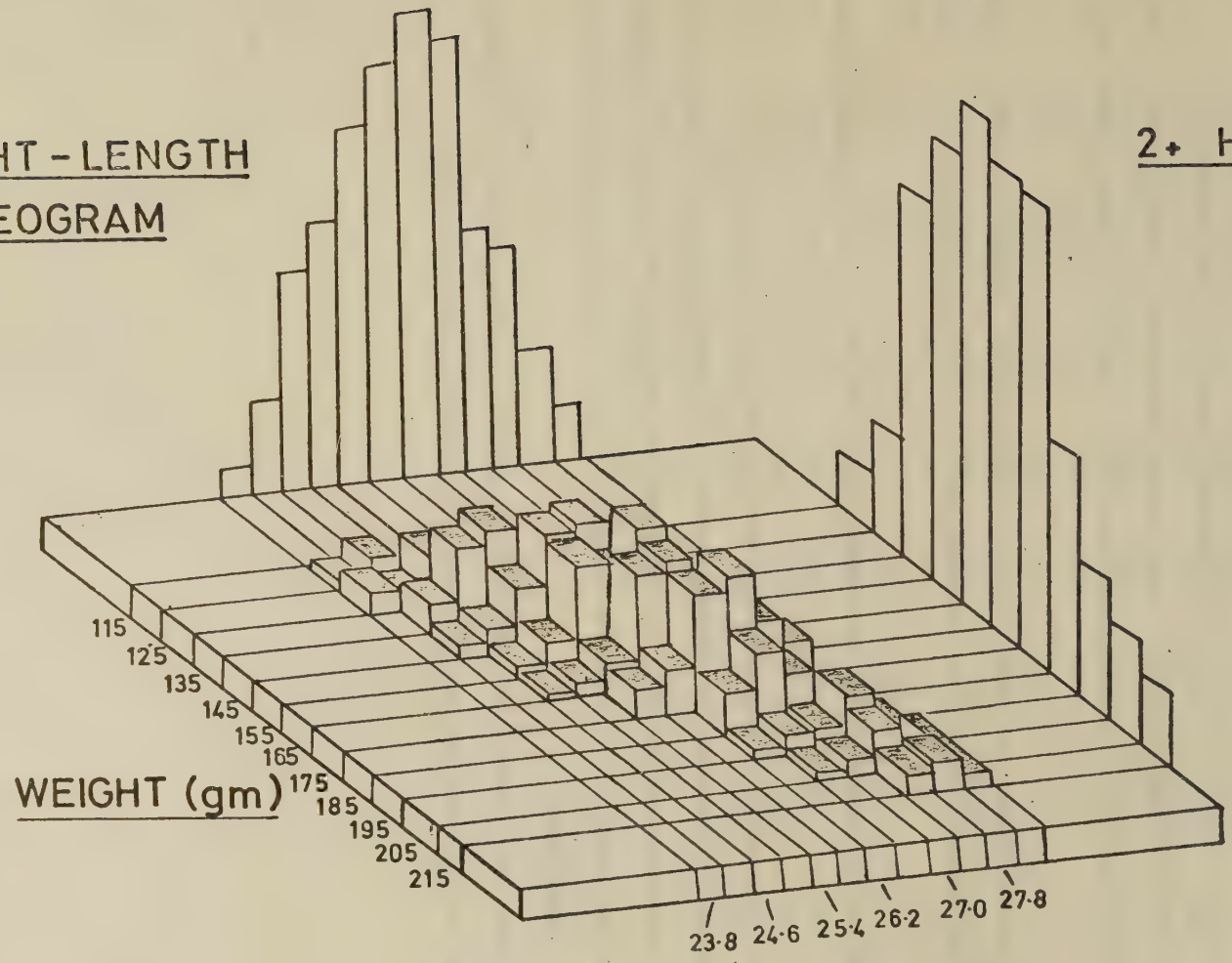

FIG.11

EXERCISES

1. The data given in Table 1 are a random sample of the weights in grams of 200 three-year old whiting arranged in 10 columns with 20 observations in each column. Arrange the observations into a frequency distribution using a suitable class interval.

Graph the resulting distribution in histogram form.

2. Using the data in Table 1, grouped in class intervals 180-189, 190-199, etc. calculate (a) the median weight, (b) the average weight and (c) the standard deviation of the weights.

3. Regarding the 200 observations in Table 1 as a random sample from a large population, calculate (a) the 99% confidence limits for a single observation and (b) the 95% confidence limits for the true mean weight.

4. Fit a Normal distribution to the data as grouped in exercise 2. Using the χ^2 test of goodness of fit show that the frequency distribution of these data is adequately represented by a Normal distribution.

5(a). What proportion of weights in the whole population is expected to lie between 215 gm and 270 gm?

(b). Above what value are 90% of the weights in the population expected to lie?

6. A random sample of 20 three-year old whiting, taken in a different area from those referred to in Table 1, gave a mean weight of 243 gm. Assuming that the true standard deviation of whiting weights is 20.5 gm, test whether the difference between this mean and that of the 200 observations given in Table 1 is significant.

7. The data in Table 1 are listed in 20 rows of 10 observations each. Estimate the standard deviation of whiting weights from the ranges of the observations in any five rows.

Table 1. Weight in grams of 200 three year old whiting

226	242	202	228	204	236	222	236	244	242
246	212	228	236	208	256	225	205	276	250
201	246	214	256	246	257	256	260	239	231
217	261	232	210	225	233	243	246	263	222
212	235	250	223	245	231	220	258	220	198
275	232	266	237	235	269	256	229	235	254
227	205	225	254	214	250	222	254	250	228
218	246	237	260	244	230	254	243	245	186
227	221	288	235	191	223	268	207	234	212
234	231	215	241	255	269	195	245	255	254
236	254	260	212	202	201	220	218	260	223
192	219	242	182	253	264	237	244	248	203
244	233	220	217	234	188	237	253	219	209
278	236	228	222	249	202	230	247	210	241
221	232	238	221	236	243	222	213	244	276
256	226	216	248	245	226	286	239	206	229
234	264	211	251	245	252	255	244	216	262
212	226	215	240	241	198	265	203	215	242
212	282	215	215	222	239	228	235	259	238
217	204	234	228	247	245	238	231	244	262

ANSWERS

1.

Class interval	Mid point	Frequency
180-189	185	3
190-199	195	5
200-209	205	15
210-219	215	26
220-229	225	33
230-239	235	37
240-249	245	34
250-259	255	25
260-269	265	15
270-279	275	4
280-289	285	3

2.

a) 235

b) 234.75

c) $\pm$ 20.48

3.

a) $234.75 \pm 2.58(20.48) = 181.91$ to 287.59

b) $234.75 \pm \frac{1.96(20.48)}{\sqrt{200}} = 231.91$ to 237.59

4.

Weights	185	195	205	215	225	235	245	255	265	275	285
Observed frequency	3	5	15	26	33	37	34	25	15	4	3
Expected frequency	2.8	6.2	13.6	24.6	34.6	38.8	33.4	24.2	13.2	5.8	2.8

$\chi^2 = 1.483$ (degrees of freedom = 8)

5.

a) 0.79

b) 209 gm

6.

$$N.D. = \frac{243-234.75}{20.5\sqrt{\left(\frac{1}{200}+\frac{1}{20}\right)}} = 1.72 \text{ (not significant)}$$

FAO FISHERIES TECHNICAL PAPER (MFS)

Documents in this group are provisional editions of volumes or sections of FAO MANUALS IN FISHERIES SCIENCE, usually in the language of the original draft. They are given a limited distribution for comment by collaborators in this project, and for use of FAO headquarters and field staff and at training centers, courses and seminars. Revised versions are subsequently published by FAO in the official languages of the Organization.

FT/T41	Manual of methods for fish stock assessment - Part III. Selectivity of fishing gear	January 1966
FIb/T51 Suppl. 1	Manual of methods for fish stock assessment - Part IV - Marking	
	Manuel sur les méthodes d'évaluation des stocks ichtyologiques - IV[e] partie - Le marquage	
	Manual de métodos para la evaluación de los stocks de peces - Parte IV - Marcado	April 1966
FRs/T38 Rev. 1	Manual of methods for fish stock assessment - Part II - Tables of yield functions	
	Manuel sur les méthodes d'évaluation des stocks ichtyologiques - II[e] partie - Tables des fonctions de rendement	
	Manual de métodos para la evaluación de los stocks de peces - Parte II - Tablas de funciones de rendimiento	September 1966
FRs/T40 (Rev. 2)	Manual of methods for fish stock assessment - Part I. Fish population analysis	June 1968
FRs/T26 Suppl. 2	Manual of sampling and statistical methods for fisheries biology. Part II - Statistical methods	July 1968

A printed English version of "Manual of Field Methods in Fisheries Biology. Provisional Edition" was issued in 1960, as a revision of:

FB/58/T1	A Manual of Field Methods in Fisheries Biology (Preliminary draft)	September 1958

Drafts of "Manual of Laboratory Methods in Fisheries Biology" were issued as:

FB/58/T3	Part I - Physical and Chemical Methods in Meteorology and Oceanography	January 1958
	Part II - Laboratory Methods on Plankton and Benthos	February 1958

Revised versions of these have been combined with the Field Methods Manual to produce a "Manual of Methods in Fisheries Biology," the definitive text of which has been published in three language versions as **FAO Man. Fish. Sci.,** (1). A provisional English edition of **FAO Man. Fish. Sci.,** (2) "Fisheries Science" was issued for FAO by CSIRO, Cronulla, Australia, in 1960.

MI/76430/8.68/E/1/600